Proposal

1. Abstract

In this project, we conduct a classifier model with various methods for speech emotion recognition(SER) problems. We use data from Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Based on recent research, the key point during feature engineering is to find the essential feature that distinguish different emotions at a maximum level. After getting the essential feature metrics, such as pitch, Log-Mel Spectrogram or amplitude, we would use SVM, LSTM or CNN to train our model and try to get a better result.

2. Brief Idea about Implementation

Three key issues need to be addressed for successful SER system, namely, (1) choice of a good emotional speech database, (2) extracting effective features, and (3) designing reliable classifiers using machine learning algorithms. In fact, the emotional feature extraction is a main issue in the SER system. Many researchers have proposed important speech features which contain emotion information, such as energy, pitch, formant frequency, Linear Prediction Cepstrum Coefficients (LPCC), Mel-frequency cepstrum coefficients (MFCC), and modulation spectral features (MSFs). Thus, most researchers prefer to use combining feature set that is composed of many kinds of features containing more emotional information.

However, using a combining feature set may give rise to high dimension and redundancy of speech features; thereby, it makes the learning process complicated for most machine learning algorithms and increases the likelihood of overfitting. Therefore, feature selection is indispensable to reduce the dimensions redundancy of features.

3. Dataset

We are using Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) as our dataset.
Source Link: RAVDESS Facial Landmark Tracking | Zenodo

3.1 Description

The RAVDESS is a validated multimodal database of emotional speech and song. The database is gender balanced consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity, with an additional neutral expression. All conditions are available in face-and-voice, face-only, and voice-only formats. The set of 7356 recordings were each rated 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals who were characteristic of untrained research participants from North America. A further set of 72 participants provided test-retest data. High levels of emotional validity and test-retest intrarater reliability were reported. Corrected accuracy and composite "goodness" measures are presented to assist researchers in the selection of stimuli.

3.2 Audio-only Files

Audio-only files of all actors (01-24) are available as two separate zip files (~200 MB each):

- Speech file (Audio*Speech*Actors_01-24.zip, 215 MB) contains 1440 files: 60 trials per actor x 24 actors = 1440.
- Song file (Audio*Song*Actors_01-24.zip, 198 MB) contains 1012 files: 44 trials per actor x 23 actors = 1012

3.3 Audio and Speech Files

Video files are provided as separate zip downloads for each actor (01-24, ~500 MB each), and are split into separate speech and song downloads:

- Speech files (Video*Speech*Actor*01.zip to Video*Speech_Actor_24.zip) collectively contains 2880 files: 60 trials per actor x 2 modalities (AV, VO) x 24 actors = 2880.
- Song files (Video*Song*Actor*01.zip to Video*Song_Actor_24.zip) collectively contains 2024 files: 44 trials per actor x 2 modalities (AV, VO) x 23 actors = 2024

3.4 File Summary

Each of the 7356 RAVDESS files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 02-01-06-01-02-01-12.mp4). These identifiers define the stimulus characteristics:

Filename identifiers

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Repetition (01 = 1st repetition, 02 = 2nd repetition).

- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

Filename example: 02-01-06-01-02-01-12.mp4

1. Video-only (02)
2. Speech (01)
3. Fearful (06)
4. Normal intensity (01)
5. Statement “dogs” (02)
6. 1st Repetition (01)
7. 12th Actor (12)
8. Female, as the actor ID number is even.

4. Reference

[1] Emotion-Recognition/Emotion Recognition from Speech- Project Report.pdf at master · rajamohanharesh/Emotion-Recognition · GitHub

[2] The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) | Zenodo

[3] https://www.ijrte.org/wp-content/uploads/papers/v7i4s/E1917017519.pdf

[4] https://arxiv.org/pdf/1906.05681.pdf

[5]https://arxiv.org/pdf/1904.06022v1.pdf

[6]https://arxiv.org/pdf/1810.04635v1.pdf